Image Retrieval: The MIRACLE Approach

Julio Villena[1,2], José L. Martínez[1,3], Jorge Fombella[1], Ana G. Serrano[4], Alberto Ruiz[4], Paloma Martínez[3], José M. Goñi[5], and José C. González[1,5]

[1] DAEDALUS, S.A.
Centro de Empresas "La Arboleda", Ctra. N-III, km 7,300 – 28031 Madrid, Spain
{jvillena,jfombella,jgonzalez}@daedalus.es
[2] Department of Telematic Engineering, Universidad Carlos III de Madrid,
Avda. Universidad 30 – 28911 Leganés, Madrid, Spain
jvillena@it.uc3m.es
[3] Computer Science Department, Universidad Carlos III de Madrid,
Avda. Universidad 30 – 28911 Leganés, Madrid, Spain
{jlmferna,pmf}@inf.uc3m.es
[4] Artificial Intelligence Department, Universidad Politécnica de Madrid,
Campus de Montegancedo s/n – 28660 Boadilla del Monte, Spain
{agarcia,aruiz}@isys.dia.fi.upm.es
[5] E.T.S.I Telecomunicación, Universidad Politécnica de Madrid,
Ciudad Universitaria s/n – 28040 Madrid, Spain
jmg@mat.upm.es, jgonzalez@dit.upm.es

Abstract. ImageCLEF is a pilot experiment run at CLEF 2003 for cross language image retrieval using textual captions related to image contents. In this paper, we describe the participation of the MIRACLE research team (Multilingual Information RetrievAl at CLEF), detailing the different experiments and discussing their preliminary results.

1 Introduction

There are two different approaches for image retrieval: content-based and text-based. Although during the last few years great efforts have been made in content-based image retrieval, it is commonly accepted that, up to now, the current state-of-the-art cannot solve the retrieval problem satisfactorily. Thus, we are focusing on text-based image retrieval, where the idea is to associate a text description with each image that describes its visual contents, and use it for the retrieval process. Cross Language Image Retrieval (CLIR) is the particular case where user queries are expressed in a language different to that of the image descriptions.

Image retrieval has its own characteristics that make it different from general text (or document) retrieval [1]. Image descriptions are usually incomplete, only showing partial aspects of the whole visual content and thus limiting the search options, and tend to be fairly short (typically image captions and/or a few keywords referring the most relevant characteristics of the image). User queries are generally more specific in image retrieval than in text retrieval [12] (users often look for images containing

C. Peters et al. (Eds.): CLEF 2003, LNCS 3237, pp. 621–630, 2004.

specific contents –e.g., "fisherman in a boat"– instead of general categories –"boats"), and are even shorter than image descriptions (typically two or three words).

ImageCLEF [10] is a pilot experiment run at CLEF 2003 [11], which consists of cross language image retrieval using textual captions. A collection of nearly 30,000 black and white images from the Eurovision St Andrews Photographic Collection [10] was provided by the task coordinators. Each image had an English caption (of about 50 words). Sets of 50 topics in English, French, German, Italian, Spanish and Dutch were also provided. Non-English topics were obtained as a human translations of the original English ones, which also included a narrative explanation of what should be considered relevant for each image.

The proposed experiments were designed to retrieve the relevant images of the collection using different query languages, therefore having to deal with monolingual and bilingual image retrieval (multilingual retrieval is not possible as the document collection is only in one language). Although there are clear limitations in the current ImageCLEF task, both in the size of the collection and the number of possible experiments to be carried out (six – one monolingual and five bilingual), it represents an interesting starting point to get an idea of the performance of CLIR systems, both in monolingual and bilingual searches, and promote research into this information retrieval field.

The MIRACLE (Multilingual Information RetrievAl at CLEF) team is a joint effort of different research groups from two universities and one private company, with a strong common interest in all aspects of information retrieval and a long-lasting cooperation in numerous projects. In this paper we describe the different experiments that were submitted to the ImageCLEF 2003 campaign. The techniques applied vary from automatic machine translation, strategies for query construction, relevance feedback to topic term semantic expansion using WordNet [6]. The main objective behind the MIRACLE participation is to compare how these different retrieval techniques affect retrieval performance.

2 Description of the MIRACLE Experiments

The MIRACLE team submitted 25 runs to ImageCLEF, based on different system parameters: 5 for the monolingual English task, 6 for the bilingual Spanish to English and German to English tasks and 4 for the bilingual French to English and Italian to English tasks. All submitted runs are automatic (no human intervention in the whole retrieval process). As previously stated, all experiments are based on text-based image retrieval and make use of the image captions only.

This section contains a description of the tools, techniques and experiments that have been used for the different tasks.

The core information retrieval engine was Xapian [5], which is a free software/open source information retrieval library, released under the GPL and based on the probabilistic information retrieval model [1] [2]. We chose Xapian because it is designed to be a highly adaptable toolkit to allow developers to add advanced indexing and search facilities easily to their own applications. It integrates Snowball stemming algorithms [7] (based on the Porter algorithm [8]), and its complete

implementation of the probabilistic information retrieval model allows term weighting and relevance feedback to be carried out.

In order to apply natural language processing to image descriptions and topics, ad-hoc tokenizers have been developed for each included language. They are used to identify different kinds of alphanumerical tokens such as dates, proper nouns, acronyms, etc., as well as recognising some of the common compound words from each language. Standard stopwords lists have also been used and a special word decompounding module for German has been applied. For English monolingual runs, (English) WordNet [6] has been used to expand queries with their synonyms.

Finally, for translation purposes, two available translation tools were considered: Free Translation Internet engine [3] for full text translations, and ERGANE dictionary lookup [4] for word by word translations.

At an initial stage common to all experiments, Xapian was used to index all the image descriptions in a single database. For each image, only the HEADLINE and TEXT fields were considered to create the image description, which was then tokenized, stemmed and stopword filtered with the English modules, before indexing it with Xapian.

We wanted our experiments to address the query construction and result merging issues. All of the previous modules were coupled in different ways, in order to evaluate different approaches for creating the query from the topic and to compare the influence of each one on the precision and recall of the image retrieval process. The name of each experiment reflects the techniques that were used in each case and the languages of the topics and the collection (always English).

2.1 Monolingual Experiments

In all cases, both the topic and the document language was English ("en"). Each of the 5 runs submitted consisted in one of the following base experiments (Q="query"):

- **Qor:** Intended as the baseline experiment to be compared with the results of other experiments, it consists of building the query with the combination of all the stemmed words appearing in the TITLE topic field, without stopwords, using an OR operator between them and including term weighting (the relative frequency of appearance of the stem in the topic).
- **Qorlem:** This experiment uses both the original words of the topic and the stemmed words, using the same OR operator and term weighting as before, i.e., it resembles the previous experiment but adds the original (non-stemmed) word forms. The idea behind this experiment is to try and measure the effect of inadequate word stemming.
- **Qorlemexp:** The idea behind this experiment is to perform synonym expansion of the terms and stems used in the previous Qorlem experiment, linking the newly obtained words with an OR operator, with the objective to retrieve a larger documents set (increase recall), despite a reduction in precision.
- **Qdoc:** For this experiment, a special feature of the Xapian system was used, which allowed the carrying out of queries based on documents in contrast to the indexed document collections. The query was first indexed as if it were another

image description, and then "similar documents" to this one were retrieved as results. This approach is similar to the idea of the Vector Space Model [1].

- **Qorrf:** This experiment carries out a blind relevance feedback (based on the results of a simple OR query as in the Qor experiment). The process consists of creating a query, getting the first 25 documents, extracting the 250 most important terms for those documents (top 10 keywords of each one), and constructing a new query to be carried out against the index database, which would provide the final results.

2.2 Bilingual Experiments

In all cases, the document language was English ("en"), but the topic language ranged from Spanish ("es"), German ("ge"), and French ("fr") to Italian ("it"). 20 different runs were submitted, consisting of the combination of the following base experiments with different languages (QT="query translation"):

- **QTor1:** Similar to the monolingual Qor experiment, but using the FreeTranslation tool: first, translate the full query from the source language to English with FreeTranslation, then apply the tokenizer to identify the different tokens in English, extract the stems, remove stopwords (in this case, stopstems) and finally generate a weighted-OR query with the resulting terms, as in the monolingual Qor experiment.
- **QTor3:** In this case, in addition to the translation of the complete query, a word by word translation is added, using the ERGANE dictionary lookup. The other steps (tokenizing, stemming and filtering) are the same as in the QTor1 experiment. The idea is to try to improve retrieval performance by putting together different translations for the words in the query.
- **QTdoc:** This is the bilingual equivalent of the monolingual Qdoc experiment. This time the query is first translated using FreeTranslation and the result obtained is indexed by the system as if it were just another image description. The information retrieval engine (Xapian) is then asked to retrieve similar documents to this newly added one.
- **QTor3exp:** This is the bilingual equivalent of the monolingual Qorlemexp experiment. It is basically the same as the QTor3 experiment, but adding a synonym expansion (using Wordnet) of the translated terms.
- **QTor3full:** Similar to the QTor3 experiment, but adding the original query (in the original language) to the terms used in the OR query. This way, query terms incorrectly translated or that have no proper translation into English are included in their original form (possibly being of little interest, but at least appearing).
- **TQor3fullexp:** This experiment is a combination of QTor3full and QTor3exp, using both translation engines together with the original query, adding synonym expansion for all the terms obtained.

All of these experiments were submitted for the bilingual Spanish to English and German to English tasks. For the bilingual French to English and Italian to English tasks, the semantic expansion was not included as a result of time limitations.

3 Evaluation of Results

To assess the defined experiments [10], the CLEF evaluation staff used the first 100 results of each submission (45 in all) to make a document pool (different for each query). In addition, the results of manually interactive searches were also added to each pool. Then, two different assessors evaluated all of the documents in the pools, taking into account a ternary scale: *relevant*, *partially relevant* and *not relevant*. The partially relevant judgment was used to pick up images which the judges thought were in some way relevant, but could not be entirely confident.

As a final step, four relevance sets were created using the relevance judgments of both judges: *union-strict* (the images of this set were the union of the ones judged as relevant by any assessor), *union-relaxed* (the union of the images judged as relevant or partially relevant by any assessor), *intersection-strict* (images judged as relevant by both assessors) and *intersection-relaxed* (images judged as relevant or partially relevant by both assessors). Strict relevance and intersection sets can be considered as high-precision results, while relaxed relevance and union sets can be thought of as results which promote higher recall.

In this section, we will present the results obtained in our experiments to reach some conclusions relative to the different approaches.

3.1 Monolingual Task

As stated before, the monolingual task consists of a set of queries in English, derived from a collection of image descriptions also in English. Figure 1 shows the recall vs. precision graph for each of the five runs we carried out for this task. The values presented correspond to the evaluation of the results, comparing them with the *intersection-strict* relevance set (the more stringent one).

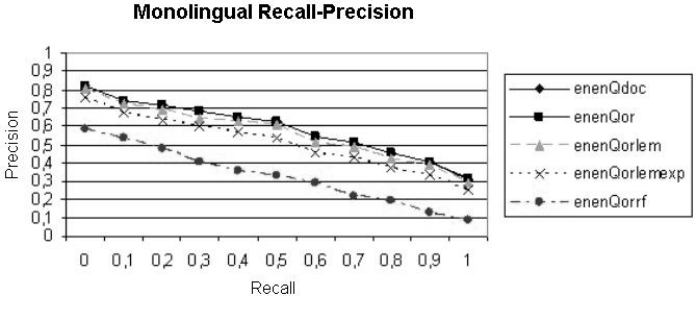

Fig. 1. Recall-Precision graph for the Monolingual task

The figure shows that the best runs have a fairly high precision value, specially taking into account that image retrieval is a difficult task. In fact, the results appear to be too high if we compare them with the monolingual document retrieval results that we obtained in the CLEF 2003 [9] monolingual tasks. Our interpretation is that the

actual coverage of relevant documents was not as complete as it should have been, because of the way the relevant sets were established (based on the submissions of every group) and because only four groups took part in ImageCLEF this year. That could be the reason why such high precision values have been obtained.

The run using blind relevance feedback leads to considerably worse results than all the other strategies. A possible explanation could be that the parameter values used in the automatic relevance feedback were not appropriate to the kind of documents we were trying to retrieve. In fact, we used the top 250 terms from the first 25 images retrieved. Given that each image has a mean description field length of 50 words, it becomes quite apparent that the number of relevant terms retrieved could be excessive. Therefore, instead of helping to locate more relevant images, these terms only add noise that seriously diminishes the overall performance.

It is worth mentioning that, instead of increasing the performance of the system, using any kind of term expansion (adding original words from the topic or performing synonym expansion) only reduces the precision of the results. This could be due to the relatively low number of images in the collection, which would not make it necessary to use term expansion to minimize the effect of heterogeneous descriptions that would arise in larger collections from different sources. Perhaps this strategy could be of interest in next ImageCLEF track, which, probably, will include larger collections.

Figure 2 represents the average precision of each submitted run for all of the topics, ordered from best to worst. This graph is a simpler representation of the overall performance value for each experiment, allowing to compare the quantitative differences of each approach. It clearly shows the poor performance of our relevance feedback experiment, and the similarity of the other experiments, especially the simple weighted-OR query approach (Qor) and the query-indexing approach (Qdoc).

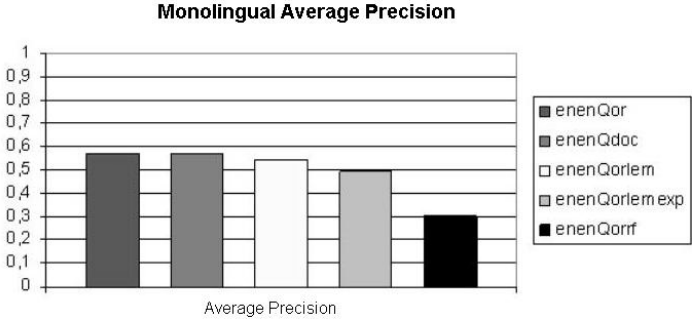

Fig. 2. Precision comparison of different runs

Although only *intersection-strict* relevance sets have been mentioned in this section, differences with the others are subtle, apart from a slight increase in the overall precision in all cases due to the larger number of relevant documents.

3.2 Bilingual Tasks

The bilingual tasks consist of the processing of queries in languages other than English, trying to retrieve relevant documents from a set of images described in English. Although queries in Spanish, Italian, German, French and Dutch were available, we only took part in tasks for the first four languages. Figure 3 shows the precision vs. recall graphs obtained for each of the runs carried out and for the language pairs (evaluating with *intersection-strict* relevance set).

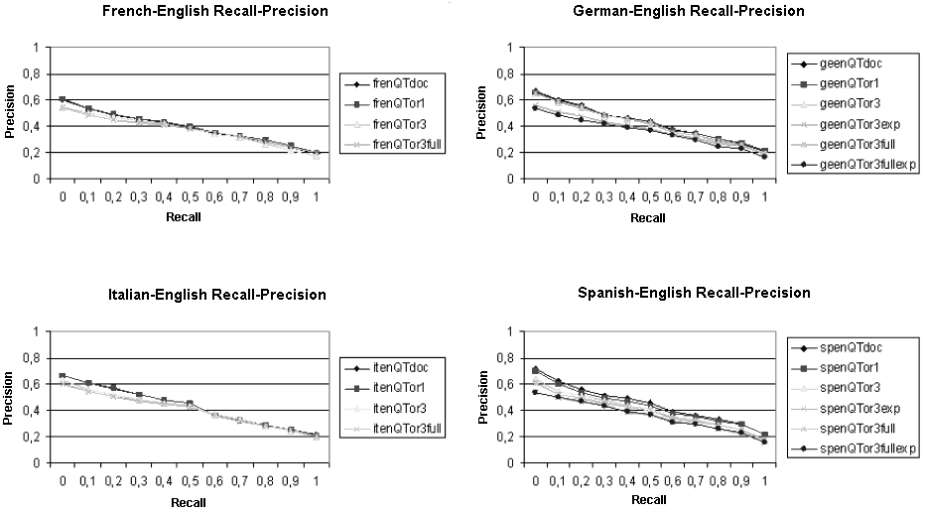

Fig. 3. Recall - Precision graphs for bilingual tasks

Several conclusions can be drawn from these figures. The most remarkable one could be the similarity between QTdoc, QTor1, QTor3 and QTor3full experiments. QTor1 and QTdoc were the best in all cases. This is somehow consistent with the results obtained in the monolingual task, where the best performance was obtained by simple OR-ing the topic terms (enenQor), and by indexing the query as another image description and searching for similar documents in the system (enenQdoc).

Another interesting aspect is that the use of more than one automatic translation has shown to be worse in our case than just using one of some quality (as the FreeTranslation has proved to be). The use of ERGANE as the word by word translator should be studied in more detail to see if it was the cause of this loss of quality (bad translations or incorporation of ambiguity of meanings) or whether this quality loss was due to the new values for the term weights modified after the inclusion of word by word translation. Our impression is that the longer the query, the worse the precision (but the better the recall, we hope). An example can be found in German to English and Spanish to English runs, in which synonym expansion is included (longer queries), leading, as expected, to worse precision values.

That precision values obtained in each task are quite similar, except for the French to English queries, which were slightly worse than the others. The explanation for this could be the poorer French to English translations provided by FreeTranslation, or the use of different terms (hardest to translate) in the French queries.

Figure 4 shows the average precision of every run, in descending order of precision and grouped by tasks. As in the case of the monolingual task, the results show little difference between the different approaches, although consistently outperformed the others. It is once more apparent that our French to English retrieval results are slightly worse than the others, while the Spanish to English has obtained the best individual results (while not the best average results in all runs).

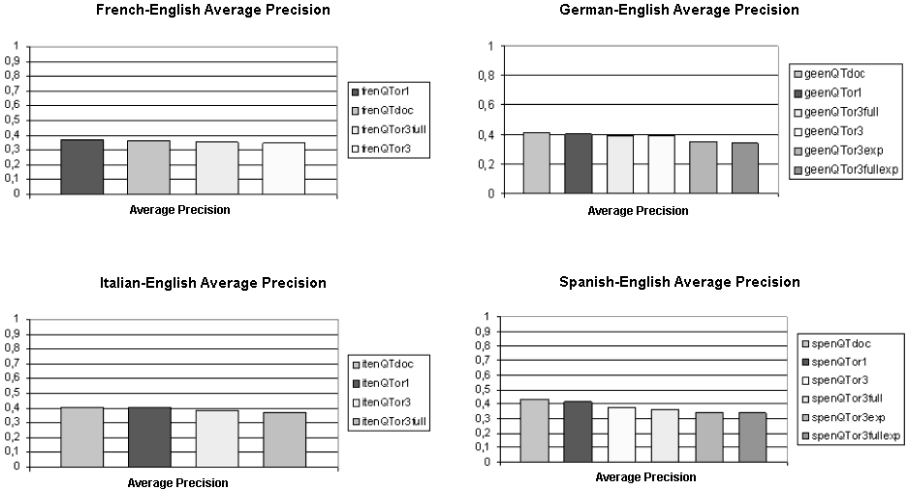

Fig. 4. Precision comparison between runs

3.3 Comparison with Other Participants

Three other groups participated in ImageCLEF 2003: the University of Surrey, the National Taiwan University (NTU), and the University of Sheffield, as the task coordinators. NTU translated the topic titles into Chinese and submitted runs for Chinese to English only, thus no comparison is possible. Although the University of Surrey submitted runs for each language, due to a misconfiguration problem with their system, the submitted results were not correct. Therefore, comparison is only fair between Sheffield and MIRACLE. The final results are shown in Table 1.

While MIRACLE obtained the best precision values in English (monolingual) and Spanish to English tasks, Sheffield exceeded our results in German and French to English. In Italian to English, the bilingual task results of the two groups were very similar.

Comparing the overall performance of the bilingual tasks with the monolingual one, there is a difference of about 10 to 15%, which is quite normal in typical CLIR

nowadays. This is aligned with similar values that we have obtained in bilingual tasks of the CLEF 2003 core track [9] (as could be expected).

Table 1. Best Mean Average Precision values for each language and group

Source Language	Sheffield	MIRACLE
English	0.5616	0.5718 (Qor)
Italian	0.4047	0.4043 (QTdoc)
German	0.4285	0.4083 (QTdoc)
French	0.4380	0.3710 (QTor1)
Spanish	0.4076	0.4323 (QTdoc)

4 Conclusions and Future Directions

The main conclusion that can be extracted from the results obtained is that the simplest approaches studied (weighted-OR-ing terms and indexing the query and then looking for similar documents) are the ones which lead to better results.

Our main goal with this first participation in the ImageCLEF task was to establish a starting point for future research work in cross-language information retrieval applied to image (and in general other non-textual types of data that can be represented somehow by textual descriptions, such as video). From our results, it is clear that there is much room for improvement both in monolingual and bilingual retrieval performance.

Also, despite the apparent poor results derived from performing synonym expansion, for us it still seems to be an interesting field of research, especially for its application to wider and more heterogeneous collections.

Acknowledgements

This work has been partially supported by the projects OmniPaper (European Union, 5th Framework Programme for Research and Technological Development, IST-2001-32174) and MIRACLE (Regional Government of Madrid, Regional Plan for Research, 07T/0055/2003).

We would like to strongly thank Paul Clough for all of his support and encouragement. Furthermore, we acknowledge the great work of both the CLEF organization team, specially Carol Peters, and Sheffield University, for the coordination of ImageCLEF task.

References

1. Baeza-Yates, R., Ribeiro-Prieto B.: Modern Information Retrieval. Addison Wesley (1999).

2. SparckJones, K., Willet, P.: Readings in Information Retrieval, Morgan Kaufmann Publishers, Inc. San Francisco, California (1997).
3. Free Translation, http://www.freetranslation.com.
4. Ergane Translation Dictionaries, http://dictionaries.travlang.com.
5. The Xapian Project, http://www.sourceforge.net.
6. Miller, G.A.: WordNet: A Lexical Database for English. Communications of the ACM, 38(11):39-41 (1995).
7. Snowball Stemming Algorithms. http://snowball.tartarus.org/.
8. The Porter Stemming Algorithm. http://www.tartarus.org/~martin/PorterStemmer/.
9. Martínez, J.L., Villena, J., Fombella, J., García Serrano, A., Ruiz, A., Martínez, P., Goñi, J.M., González, J.C.: Evaluation of MIRACLE Approach Results at CLEF 2003. Working Notes for the CLEF 2003 Workshop (21-22 August, Trondheim, Norway), Vol 1 (2003).
10. Clough, P., Sanderson, M.: The CLEF 2003 Cross Language Image Retrieval Task. Working Notes for the CLEF 2003 Workshop (21-22 August, Trondheim, Norway), Vol 1 (2003).
11. Peters, C. Introduction. Working Notes for the CLEF 2003 Workshop (21-22 August, Trondheim, Norway), Vol 1 (2003).
12. Goodrum, A.A.: Image Information Retrieval: An Overview of Current Research. Informing Science, Vol 3(2):63-66 (2000).